2024-2 시계열자료분석팀 클린업 1주차

소속	성균관대학교 통계분석학회 P-SAT
작성자	33기 김나현

목차

I. 시계열자료분석 개요

1. 시계열 자료의 정의 및 목적

시계열 자료(time series)란 시간에 따라 관측된 자료의 집합을 의미합니다. 일반적으로 시점 t에 대하여 $\{X_t, t = 1, 2, 3, ...\}$으로 표현합니다. 이때 t가 이산형인지, 연속형인지에 따라 이산형 시계열 자료와 연속형 시계열 자료로 구분할 수 있습니다.

> 🥦 **Notation**
>
> $\{x_t, t \in T_0\}$, when t is a time index and T_0 is the set of all possble time points.
>
> i. 이산형 시계열 (discrete TS) : $\{x_t\}$ if $T_0 \in Z$ (정수)
>
> ii. 연속형 시계열 (continuous TS) : $\{x_t\}$ if $T_0 \in R$ (실수)

시계열자료분석이란 시계열 자료와 추세 분석을 다루는 통계 기법으로, 시간 순으로 정렬된 데이터에서 관계를 찾아내고 의미 있는 요약과 통계 정보를 추출하는 과정을 의미합니다.

시계열자료분석의 목적은 두 가지로 나눌 수 있습니다.

1) 예측을 위한 분석

시계열 데이터를 활용하여 미래의 값을 예측하고자 하는 분석입니다. 이러한 예측 분석은 다양한 방법론을 포함하며, 주요 기법으로는 다음과 같은 것들이 있습니다.

→ 추세분석, 평활법, 분해법, 자기회귀누적이동평균(ARIMA) 모형 등

2) 시스템을 이해하고 제어하기 위한 분석

시계열 데이터의 패턴을 통해 시스템의 동작 원리와 구조를 이해하고, 이를 바탕으로 시스템을 제어하거나 최적화하기 위한 분석입니다. 이 분석은 다음과 같은 기법을 포함합니다.

→ 스펙트럼 분석, 개입분석, 전이함수 모형 등

클린업에서는 예측, 즉 forecasting을 위한 분석법에 대해 집중적으로 다루고자 합니다.

2. 시계열 자료의 특징

시계열 자료는 시간에 따라 관측되었기 때문에 시간의 흐름이 반영되어 관측치(observation) 간의 연관성(dependency)가 존재합니다. 따라서 특정 시점에 대한 확률 변수 $\{X_t\}$의 분포는 하나의 관측치만을 고려한 것이 아닌, 전체 시점에서의 관측치 집합 $\{x_1, x_2, ...\}$을 모두 고려한 결합 분포(joint distribution)라고 할 수 있습니다. 즉, 상당수의 통계적 분포에서 가정하는 독립성 조건을 만족하지 않는다는 것입니다. 따라서, 데이터 간의 연관성을 반영할 수 있는 적절한 분석 방법이 필요하고, 그 분석법이 우리가 공부할 시계열 분석입니다.

3. 시계열 자료의 구성 요소

시계열 자료의 구성 요소는 크게 규칙 요소, 불규칙 요소로 나눌 수 있습니다. 아래 네 가지 구성 요소들을 분해하여 미래를 예측하는 것이 시계열 분석의 목적입니다.

i. 규칙 요소

a. **추세 변동 (Trend)**

- 시간의 흐름에 따라 관측치가 증가하거나 감소하는 추세를 가지는 변동
- 특별한 충격이 없는 한 지속되는 특성이 있음

b. **순환 변동 (Cycle)**

- 일정한 주기를 가지고 변화하지만 규칙적으로 발생하지 않는 변동
- 경제적, 사회적 요인에 의해 발생해 예측이 어려움
- 주기적인 변화가 있지만 계절에 의한 것이 아니며, 주기가 긴 경우의 변동

c. **계절 변동 (Seasonal Variation)**

- 규칙적인 주기를 가지고 발생하는 변동
- 주별, 월별, 계절별과 같이 특정 시간 간격을 가지고 반복하는 특징
- 환경적인 요인에 의해 발생하기 때문에 예측 및 처리에 용이함

ii. 불규칙 요소

d. **우연 변동 / 불규칙 성분**

- 무작위 원인에 의해 나타나 일정한 규칙성을 인지할 수 없는 변동

4. 시계열 분해

시계열 자료 분석의 전통적인 방법 중 하나인 분해법은 시계열이 여러 성분 또는 요인으로 구성되어 있다고 보아 이를 분해한 후 성분들을 각각 추정하여 원래의 시계열을 해석하려는 방법입니다. 시계열 자료를 비정상 부분(non-stationary part)과 정상 부분(stationary part)으로 분해할 수 있는데요, 추세(m_t)와 계절성(s_t)을 비정상 부분, 오차(Y_t)를 정상 부분이라고 합니다.

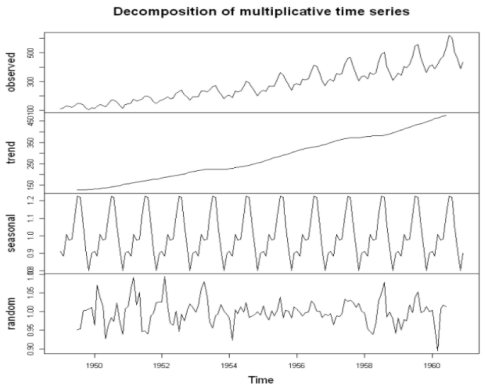
시계열 분해는 크게 덧셈 분해와 곱셈 분해로 나눌 수 있습니다.

i. **덧셈 분해(additive decomposition)**

$$X_t = m_t + s_t + Y_t$$

m_t=추세(trend), s_t=계절성(seasonality), Y_t=오차 (stationary error)

덧셈 분해는 위 식과 같이 시계열 자료를 구성 요소로 분해하는 것을 의미합니다. (원칙상으로 Y_t는 정상성을 만족해야 합니다.) 시계열 분석에서는 m_t(추세)와 s_t(계절성)을 제거한 후 정상성을 만족하는 오차를 이용해 예측 모델링을 진행합니다. 이때 추세와 계절성은 t에 대한 함수 개념으로 이해하시면 될 것 같습니다! 덧셈 분해의 과정은 아래 그림을 통해 시각적으로 확인할 수 있습니다.



ii. **곱셈 분해(multiplicative decomposition)**

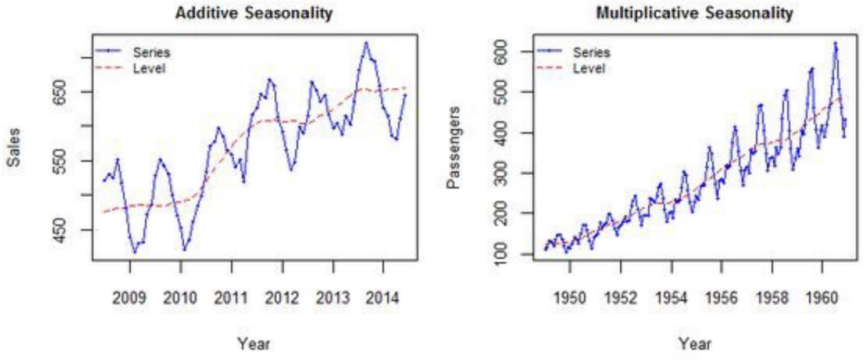
$$X_t = m_t * s_t * Y_t$$

m_t=추세(trend), s_t=계절성(seasonality), Y_t=오차 (stationary error)

위와 같이 구성 요소의 곱으로 시계열 자료를 분해하는 방법이 곱셈 분해입니다. 곱셈 분해를 사용하기 위해서는 데이터에 0이 포함되는지 확인해야 합니다. 만약 존재한다면 곱셈 분해를 사용할 수 없습니다.

덧셈 분해와 곱셈 분해의 차이점은 추세와 계절성의 관계를 통해 정의할 수 있습니다. 덧셈 분해는 추세와 계절성을 별개의 구성 요소로 보지만, 곱셈 분해는 추세에 따라 계절성이 변화함을 가정합니다.

아래 왼쪽 그래프는 추세와 계절성이 별개일 때, 오른쪽 그래프는 추세에 따라 계절성이 변화할 때를 나타낸 것입니다.



이후 자세히 다룰 예정이지만, 시계열 모형은 정상성을 가정합니다. 하지만 시간에 따라 변동 폭이 일정하지 않을 때에는 정상성을 만족하지 못하게 되어, 곱셈 분해의 식에 로그를 취해 덧셈 분해로 나타내어 계산하기도 합니다. 클린업에서는 덧셈 분해를 이용한 시계열 분석법에 대해 공부하도록 하겠습니다.

II. 정상성

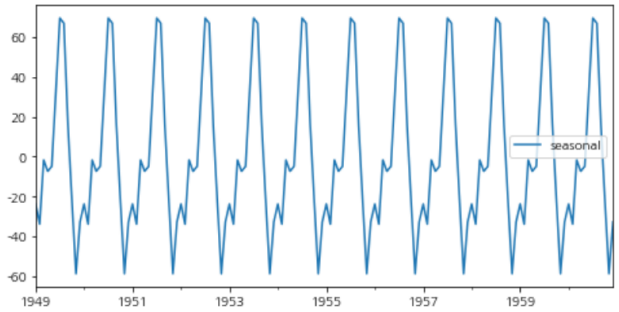
1. 정상성 가정의 필요성

정상성이 무엇인지 알아보기 전에, 정상성 가정이 왜 시계열분석에 있어 필요한지 알아보겠습니다. 미래의 값을 예측하는 목적을 지닌 시계열 모델의 경우, 추정하고자 하는 분포에 미래의 index를 포함해줘야 합니다. 즉, 시계열 자료 $X = (X_1, ..., X_n)'$의 결합분포는 $F_X(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$과 같이 표현하는데, 이때 미래의 값을 예측하기 위해서는 무한한 시점들 $X = (X_1, ..., X_n, X_{n+1}, ...)'$의 결합분포를 고려해야 한다는 것입니다.

이처럼 미래의 값을 예측하기 위해서는 무한한 시점들의 결합분포를 고려해야 합니다. 즉, 자료는 유한함에도 불구하고, 무한의 dimension에 대한 분포를 계산해야 한다는 것입니다. 그러나 이는 현실적으로 매우 복잡하기에, 몇 가지의 가정을 통해 전체 데이터를 단순화하고자 합니다. 이 가정을 **정상성 (stationarity)** 가정이라고 합니다.

2. 정상성 개념

정상성이란 **시계열의 확률적 성질이 시간에 흐름에 영향을 받지 않는 것(time-invariant)**을 의미합니다. 즉, 평균, 분산 등에 변화가 없는 것을 말합니다. 좀 더 엄밀히 정리하면, 정상성은 시계열 자료의 확률적인 성질이 시점에 의존하지 않고 시차(lag)에만 의존한다는 특성을 의미합니다. 대부분의 시계열 이론은 정상성을 가정으로 전개됩니다.



위 그림은 일반적은 시계열 그래프입니다. 이렇게 일정한 간격을 두고 반복되는 데이터라면, 즉 확률적 성질이 시차에만 의존한다면 시차 내에서의 분포를 구해 편리하게 전체 분포를 예측할 수 있습니다.

3. 강정상성

강정상성 (Strict Stationarity)

$$(X_{t_1}, ..., X_{t_n}) \overset{d}{=} (X_{t_1+h}, ..., X_{t_n+h})$$

모든 h와 n에 대하여 시계열 $\{X_t, t \in Z\}$가 위 조건을 만족할 시, 해당 시계열은 **강정상성**을 만족합니다. t_1부터 t_n까지의 n기간만큼의 시계열에 대한 결합 분포는 시점을 h만큼 옮겼을 때에도 동일한 기간에 대해서는 같은 결합 분포를 가져야 함을 의미합니다. 즉, 강정상성은 일정한 시차 간격을 가지는 관측치 집합들이 모두 같은 분포를 따르는 것을 말합니다.

그러나 여전히 분포에 대한 가정이 포함되어 이를 만족하는 것은 현실적으로 어려우며 매우 복잡합니다.

4. 약정상성

약정상성 (Weak Stationarity)

i. $E[|X_t|]^2 < \infty, \ \ \forall t \in Z$

: **2차 적률이 존재하고 시점 t에 관계 없이 일정하다.**

ii. $E[X_t] = m, \ \ \forall t \in Z$

: **평균이 상수로 시점 t에 관계없이 일정하다.**

iii. $\gamma_X(r, s) = \gamma_X(r + h, s + h), \ \ \forall r, s, h \in Z,$ or $Cov(X, X_t)$ *does not depend on* t

:

자기공분산은 시차 h에만 의존하고 시점 t와 무관하다.

시계열 $\{X_t, t \in Z\}$가 위의 세 가지 조건을 만족할 대, **약정상성**을 만족한다고 합니다. 분포 전체가 동일해야 하는 강정상성과는 달리 1차 적률($E[X_t]$)과 $\gamma(h)$만 고려하면 된다는 점에서 훨씬 간단합니다. 앞으로 다룰 정상 시계열의 경우 모두 약정상성을 가정합니다.

자기공분산(autocovariance)/자기상관(autocorrelation)이란?

공분산(covariance)와 상관계수(correlation coefficient)는 다들 들어 보셨죠? 그런데, 여기에 'auto'가 왜 붙은 것일까요? 먼저 공분산과 상관계수의 의미부터 떠올려봅시다. 공분산은 두 변수 사이의 관계, 그리고 상관계수는 이를 각 변수의 표준편차로 나누어주어 두 변수의 선형관계를 -1 에서 1 사이의 값으로 표현한 값입니다.

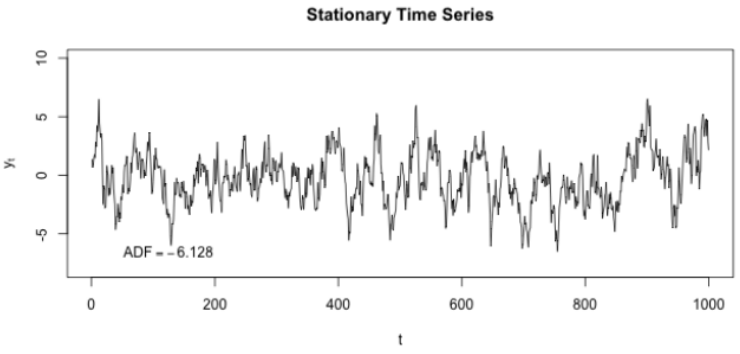
시계열에서는 두 변수가 아닌 **자기 자신과 몇 시점 떨어진 자기 자신 사이의 공분산 및 상관계수**를 구함으로써 해당 변수가 시차를 두고 어느 정도 연관되어 있는지를 확인합니다. 이제 왜 "auto"혹은 "자기"라는 말이 붙었는지 아시겠죠?

III. 정상화

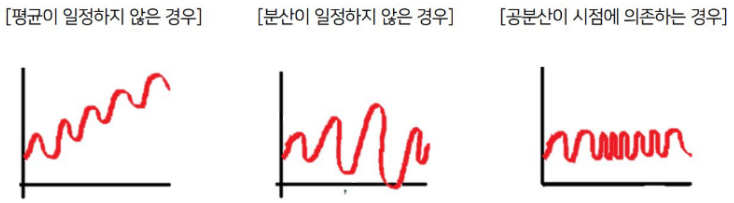
앞서 정상성 개념을 소개하며, 여러 시계열 모델이 정상성을 가정하고 전개된다고 말씀드렸습니다. 그러나 현실의 대부분의 시계열 자료의 경우 정상성을 만족하지 않습니다. 그렇다면, 이러한 비정상 시계열을 적절한 시계열 모델로 적합하기 위해선 정상 시계열로 변환해 주어야 합니다. 마치 회귀분석에서 기본 가정을 만족하지 않을 때 몇 가지 transformation 과정을 통해 맞춰주는 것처럼요! 이를 **정상화**라고 합니다.

1. 정상 시계열과 비정상 시계열

우선 주어진 시계열 자료가 정상성을 만족하는지, 아닌지를 먼저 확인해야 합니다. 이는 시계열 plot을 통해 확인할 수 있습니다. 시계열 plot의 x축은 time point t, y축은 각각의 시간에 대응하는 관측값으로 구성되어 있습니다. 이때 추세가 존재하는지, 돌발적인 변화가 있는지, 이상치가 존재하는지 등을 파악해야 합니다.



위 시계열은 정상성을 만족하는 정상 시계열로, 특별한 추세나 계절성이 보이지 않으며 평균과 분산 역시 일정한 것으로 판단할 수 있습니다.



비정상 시계열은 정상성 조건을 만족하지 못하는 시계열입니다. 위와 같이 평균 또는 분산이 일정하지 않거나, 공분산이 시점에 의존하는 시계열이 이에 해당합니다.

2. 분산이 일정하지 않은 경우의 정상화 과정

시간의 흐름에 따라 변동폭이 커지는 이분산(heteroscedasticity) 자료들이 존재합니다. 이러한 경우에 **분산 안정화 변환(Variance Stabilizing Transformation, VST)**을 통하여 분산이 시점 t에 의존하지 않고 일정하게끔 만들어주어야 합니다.

🥦 **분산 안정화 변환 방법**

i. **Box-Cox Transformation** $$f_\lambda(X_t) = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & \text{if } X_t \geq 0, \lambda \geq 0 \\ \log X_t & \text{if } \lambda = 0 \end{cases}$$

ii. **Log-Transformation** $$f(X_t) = log(X_t)$$

iii. **Square root Transformation** $$f(X_t) = \sqrt{X_t}$$

3. 평균이 일정하지 않은 경우의 정상화 과정

$$X_t = m_t + s_t + Y_t$$

m_t: 추세, s_t: 계절성, Y_t: 정상성을 만족하는 오차

위와 같이 시계열 모델을 분해했을 때, 추세가 존재하거나, 계절성이 존재하거나 혹은 추세와 계절성이 모두 존재하여 평균이 일정하지 않을 수 있습니다. 일반적으로 회귀, 평활, 차분의 3가지 방법을 통해 **비정상 부분을 추정하고 제거**하여 정상화를 진행합니다. 따라서 각각의 경우에 대하여 어떻게 정상화할 수 있는지 알아보도록 하겠습니다

i. 회귀 (Regression)

- **A [추세만 존재하는 경우] polynomial regression**

  [1] 시계열을 다음과 같이 가정합니다.

  $$X_t = m_t + Y_t,\ E(Y_t) = 0$$

  [2] 추세 성분 m_t를 다음과 같이 시간 t에 대한 선형회귀식으로 나타냅니다.

  $$m_t = c_0 + c_1 t + c_2 t^2 + ... + c_p t^p$$

  [3] 위 선형회귀식의 계수를 최소제곱법(OLS)를 통하여 추정합니다.

  $$(\hat{c_0}, ..., \hat{c_p}) = \underset{c}{\operatorname{argmin}} \sum_{t=1}^{n} (X_t - m_t)^2$$

  [4] 추정한 추세를 시계열에서 제거하면 정상 시계열이 됩니다.

- **B [계절성만 존재하는 경우] harmonic regression**

  [1] 시계열을 다음과 같이 주기가 d인 계절성만을 가진다고 가정합니다.

  $$X_t = s_t + Y_t,\ E(Y_t) = 0\ \ where\ \ \ s_{t+d} = s_t = s_{t-d}$$

  [2] 계절 성분 s_t를 다음과 같이 시간 t에 대한 회귀식으로 나타냅니다.

  $$s_t = a_0 + \sum_{j=1}^{k} (a_j cos(\lambda_j t) + b_j sin(\lambda_j t))$$

  [3] 적절한 λ_j와 k를 선택한 후, OLS를 통하여 a_j와 b_j를 추정합니다. (Appendix 참조)

  [4] 추정한 계절성을 시계열에서 제거하면 정상 시계열이 됩니다.

- **C [추세 및 계절성이 동시에 존재하는 경우]**

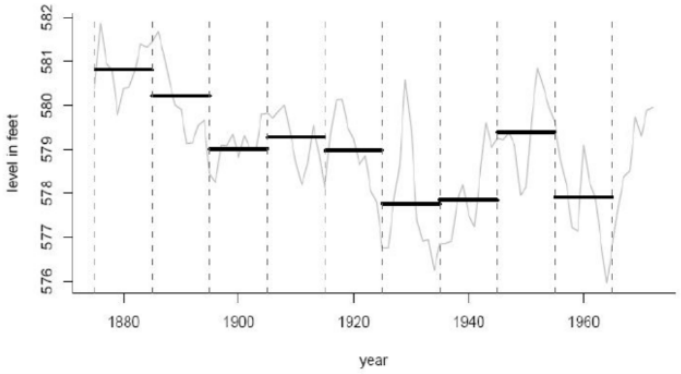
  $$X_t = m_t + s_t + Y_t,\quad E(Y_t) = 0$$

  A와 B 과정을 차례대로 진행합니다. 이후에도 남아 있는 추세가 보인다면 같은 과정을 반복해 제거합니다.

다만, 회귀 방법에 사용되는 최소제곱법의 경우 기본적으로 오차항의 독립성을 가정하고 전개되는데, 시계열의 오차항은 독립성을 가정하지는 않아 추정이 정확하지 않을 수 있습니다. 특히 분산을 계산할 때 만약 오차항이 연관되어(correlated) 있음에도 불구하고 독립을 가정한 상태로 계산되기 때문에 틀릴 수 있습니다. 이에 따라 신뢰 구간의 계산까지 오류가 생길 수 있다는 단점이 존재합니다.

ii. 평활 (Smoothing)

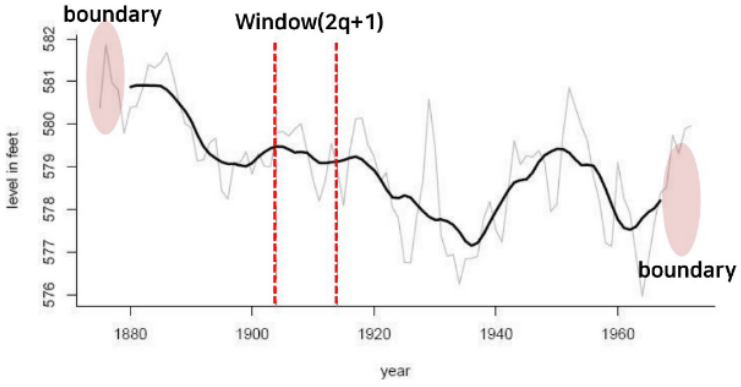
회귀는 전체 데이터를 한 번에 추정하기 때문에, 국소적 변동(local fluctuation)을 잡아내지 못 할 수 있습니다. 따라서, 국소적 변동에 주목해야 하는 경우에는 평활법을 사용할 수 있습니다.



평활법은 위와 같이 시계열 자료를 여러 구간으로 나눈 후, **구간의 평균들로 추세를 추정**하는 방법입니다. 위 그림을 통해 알 수 있는 것처럼, 전체 시계열 자료와 구간 평균의 움직임은 비슷할 것이라는 아이디어를 이용한 방법입니다.

- **A1 [추세만 존재하는 경우] 이동평균 평활법(Moving Average Smoothing)**

  이동평균 평활법은 일정기간마다 평균을 계산하여 추세를 추정하는 방식입니다.



  [1] 위와 같이 길이가 $2q+1$인 구간의 평균을 구합니다.

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{j=q} (m_{t+j} + Y_{t+j})$$

  [2] 추세가 linear하다고 가정했을 때, 위 구간의 평균 W_t는 근사적으로 추세 m_t와 같아집니다. (Appendix에 수식 참조!) 즉, 일정한 길이의 구간의 평균이 전체 시계열의 추세를 잡을 수 있다는 것이죠.

  [3] 위 과정을 통해 추세 부분만 남은 W_t를 전체 데이터에서 제거함으로써 정상 시계열을 확보할 수 있습니다.

  이동평균 평활법은 국소적 변동을 설명할 수는 있지만, 데이터의 맨 앞 q개와 맨 뒤 q개의 **bondary에 해당하는 값을 추정할 수 없게** 됩니다. 또한, 현실에서는 현재 t시점 이후의 데이터를 활용할 수 있는 경우는 거의 불가능합니다. 따라서, 과거의 데이터에만 의존하여 추정하는 지수평활법에 대해 알아보겠습니다.

- **A2 [추세만 존재하는 경우] 지수평활법(Exponential Smoothing)**

  지수평활법은 미래의 데이터를 활용하지 않고, 추세 $\hat{m}_t$를 시점 t시점까지의 관측값만을 이용하여 추정하고 제거하는 방법입니다. 즉, 과거 시점의 데이터만을 이용하여 추세를 추정한다는 점에서 이동평균평활법과 차이가 있습니다.

추세를 추정하는 방식은 다음과 같습니다. $a \in [0,1]$인 a에 대하여, **현재시점에는 a만큼, 과거의 예측값에 대해서는 $1-a$만큼의 가중치를 부여**하여 추세를 추정합니다.

$\hat{m_1} = X_1$

$\hat{m_2} = aX_2 + (1-a)\hat{m_1} = aX_2 + (1-a)X_1$

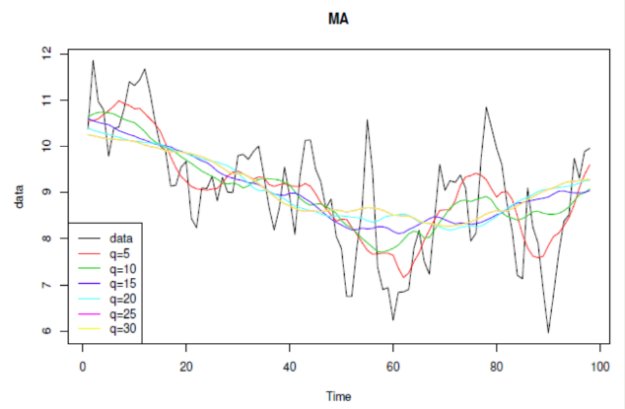
$\vdots$

$\hat{m_t} = aX_t + (1-a)\hat{m_{t-1}} = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1}X_1$

위와 같이 추정한 추세를 시계열에서 제거합니다. 이때 더 과거의 값일수록 가중치가 지수적으로 줄어드는 것을 알 수 있습니다.

평활법에서 q와 a의 선택

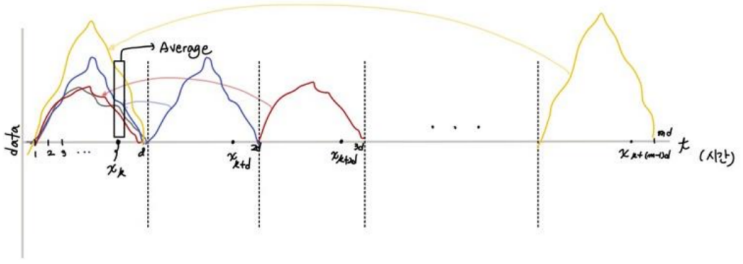
평활법은 추세 외에도 tuning parameter q와 a를 추정해야 합니다. 일반적으로 q가 작으면 작은 변화들도 잘 잡아낼 수 있지만, 그만큼 변동성이 심해집니다. 반대로 q가 클 경우 변동성은 줄어들지만 작은 변화를 잡아내지 못합니다. 즉, **bias-variance trade off** 가 발생합니다. 이러한 편향과 분산의 관계가 존재하기에 적절한 parameter를 찾는 것이 중요합니다. 일반적으로 cross-validation(CV)를 통해 MSE를 추정하여 최적의 파라미터를 선택하게 됩니다.



CV와 bias-variance trade off에 대해 더 자세히 알고 싶다면 데마팀 1주차클린업을 참고하세용

- **B [계절성만 존재하는 경우] Seasonal Smoothing**

IDEA: 주기가 d인 관측치를 한 주기 안에 모두 겹친 후 평균 내자!

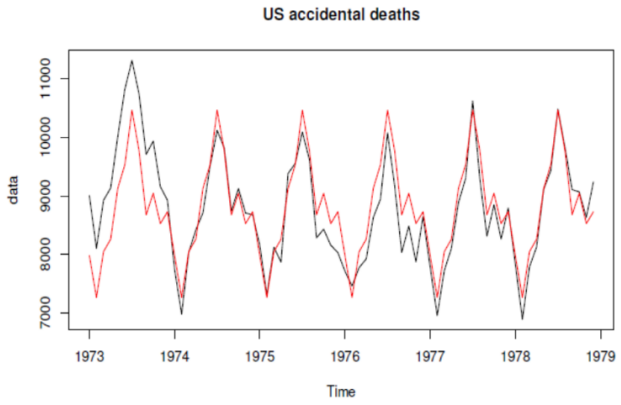


[1] $k = 1, \ldots, d$에 대하여 계절성분 $\hat{s_k}$를 추정합니다.

$$\hat{s_k} = \frac{1}{m}(x_k + x_{k+d} + \cdots + x_{k+(m-1)d}) = \frac{1}{m}\sum_{j=0}^{m-1} x_{k+jd}$$

(m=# of obs. in kth seasonal component)

[2] 추정된 계절 성분을 다른 주기에도 적용해 전체 계절성을 추정하고, 이를 시계열 자료에서 제거합니다.



위 그림처럼 빨간색으로 그려진 계절 성분은 모든 주기에서 동일하게 반복되고 있습니다.

- **C1 [추세와 계절성이 모두 존재하는 경우] Classical Decomposition Algorithm**

$$X_t = m_t + s_t + Y_t, E(Y_t) = 0$$

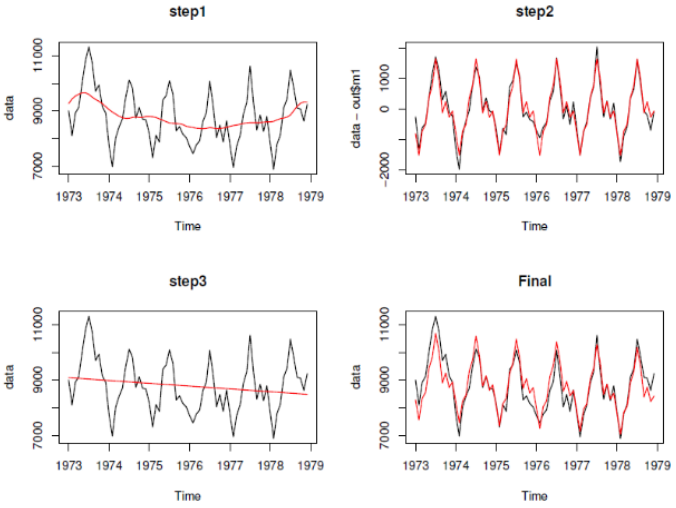
[1] MA filter를 이용하여 추세를 예측한 뒤, 추정한 추세를 제거해준 후 계절성분을 seasonal smoothing으로 추정합니다.

[2] 계절성까지 제거된 시계열에서 다시 추세를 OLS를 통해 추정합니다.

$$\hat{m_t}^{new} = argmin_c \sum_{t=2}^{n} (X_t - \hat{s_t} - c_0 - c_1 t - c_2 t^2 - \cdots - c_p t^p)^2$$

[3] 새롭게 추정한 추세와 계절성을 다음과 같이 제거해주어 오차를 추정합니다.

$$\hat{e_t} = X_t - \hat{m_t} - \hat{s_t}$$



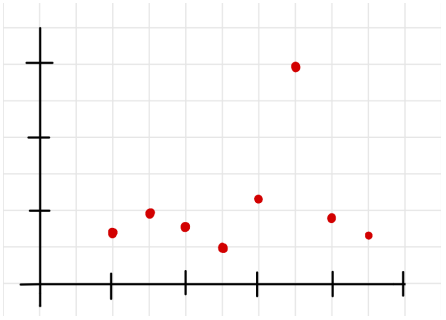
이 과정을 모두 거치고도 또 다시 추세가 있다면 다시 반복해줍니다. 다만, Seasonal Decomposition이기 때문에 초기와 마지막 일부 데이터에 대해서 추세 추정값을 알 수 없습니다. 또한 데이터에서 급격한 증가나 감소가 발생하는 부분이 있을 때 전통적 시계열 분해에서는 smoothing이 너무 강하게 일어나게 됩니다.

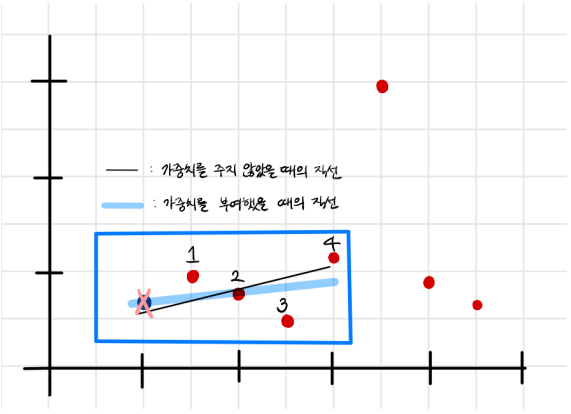
최근에는 전통적 분해법 말고도 보다 정확한 추정이 가능한 방법을 사용합니다. 가장 대표적으로, STL 분해에 대해서 알아보도록 하겠습니다.

- **C2 [추세와 계절성이 모두 존재하는 경우] Seasonal and Trend Decomposition using Loess**

STL 분해법은 Loess(Locally Weighted Scatterplot Smoothing)라는 기법을 중심으로 추세-주기와 계절성분을 구하게 됩니다.

우선, Loess 개념에 대해서 먼저 알아보도록 하겠습니다. 다음과 같은 데이터가 있다고 해보겠습니다.

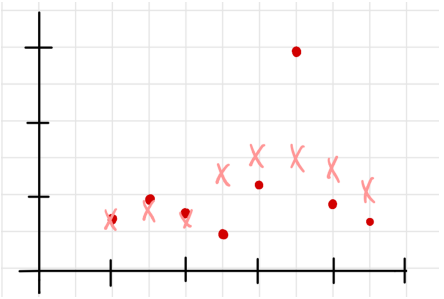




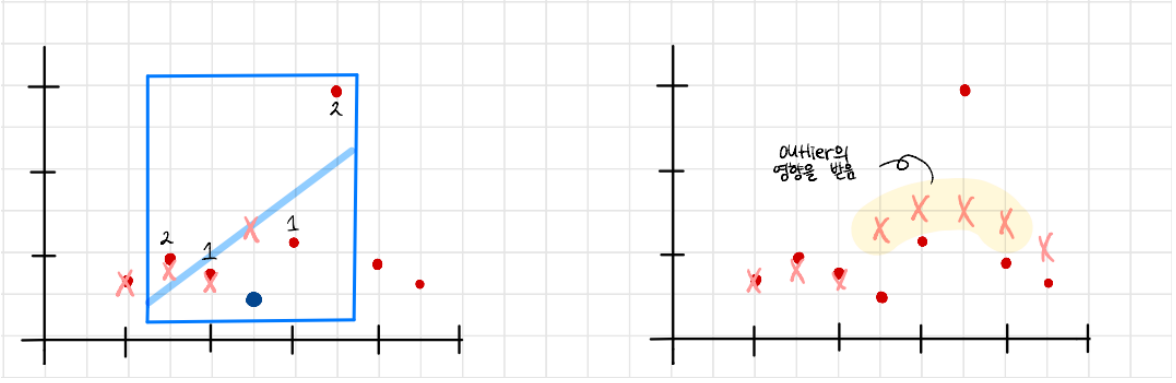
Loess란 국소적으로(특정 데이터 범위 내에서) 데이터에 가중치를 부여하여 곡선을 fitting하는 방법입니다. window size=5라고 했을 때, focal point(남색 point)에서 가장 가까운 점에 대해서 1,2,3,4번째 순서대로 큰 가중치를 부여할 수 있습니다.

이때, 가중치를 부여하지 않은 경우에 비해 x축 간의 거리를 고려한 가중치를 주었을 때의 직선이 더 잘 fitting되는 것을 확인할 수 있습니다.

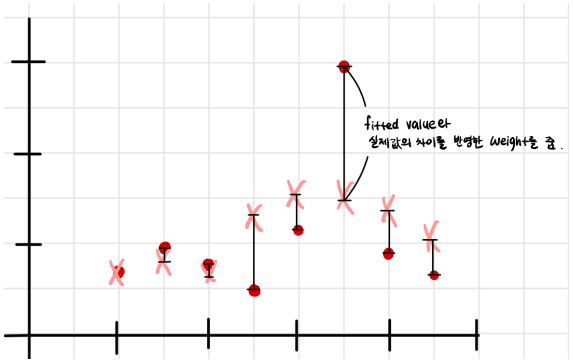
이 과정을 순차적으로 모든 데이터 포인트에 대해 진행합니다.

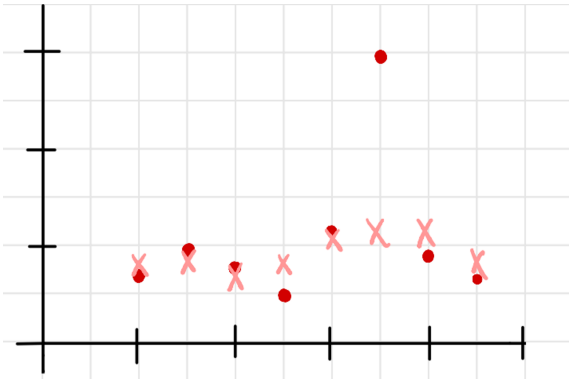


다음과 같이 outlier가 존재하는 경우, 추정된 데이터 포인트는 원래 데이터보다 훨씬 멀어지게 됩니다.

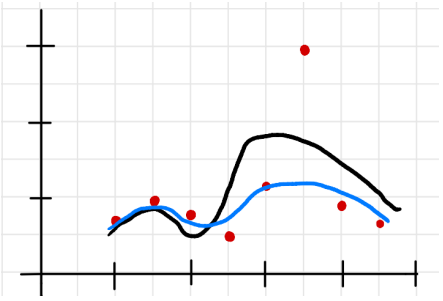


이러한 경우를 해결하기 위해, Loess는 두 번째 가중치를 부여하게 됩니다.





fitting된 값과 원래의 데이터 포인트 간의 y값이 클수록 Original data를 잘 추정하지 못했다는 것을 의미하므로 낮은 가중치를 부여하고, 첫 번째 데이터 포인트와 같이 거리가 가까울수록 높은 가중치를 부여합니다. 이 과정을 데이터의 모든 위치에 대해 반복하여 부드러운 곡선을 생성합니다.



발그림 죄송해여 ㅎ.ㅎ

위와 같이 두 단계의 가중치를 부여한다면 이상치의 영향을 덜 받음과 동시에 Original 데이터를 보다 smooth하게 적합시킬 수 있습니다. Loess는 데이터의 비선형적이고 복잡한 패턴을 포착하는 데 적합하며, 전역적 모델을 사용하지 않기 때문에 데이터의 국소적인 변화를 잘 반영할 수 있습니다.

다들 Loess의 원리가 어떻게 작동하는지 이해하셨나요? 그렇다면, 이를 활용한 STL 분해에 대해 수식적으로도 이해해보겠습니다.

$$X_t = m_t + s_t + Y_t,\ E(Y_t) = 0$$

[1] 계절 주기의 반복 평균을 이용하여 초기 계절 성분을 추정합니다.

$$\hat{S}_t = \frac{1}{K}\sum_{k=1}^{K} X_{t+kP}$$

[2] 초기 계절 성분을 더 부드럽게 하기 위해 Loess 평활화 기법을 적용합니다.

$$\hat{S}_t^{(smoothed)} = \text{Loess}(\hat{S}_t)$$

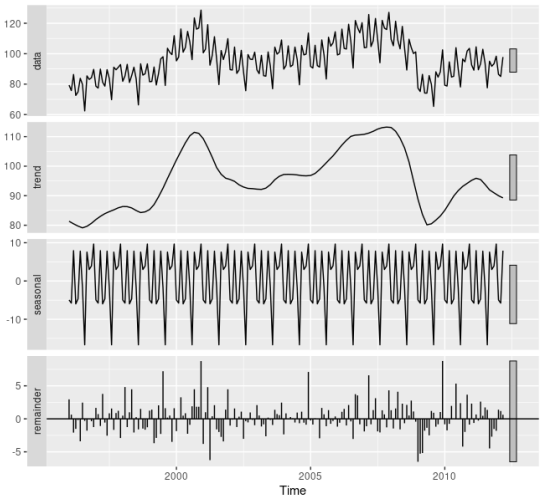
[3] 계절 성분을 제거한 후, 추세 성분 T_t를 Loess를 사용하여 추정합니다.

$$Y_t = X_t - \hat{S}_t$$

$$\hat{T}_t = \text{Loess}(R_t)$$

[4] 추정된 계절 성분과 추세 성분을 제거하여 잔차 성분을 계산합니다.

$$\hat{Y}_t = X_t - \hat{T}_t - \hat{S}_t$$



이후 계절 성분과 추세 성분의 추정을 반복적으로 업데이트합니다. 이처럼 STL 분해는 비선형 추세와 비선형 계절성을 포함하는 복잡한 시계열 데이터를 분석하는 데 유용합니다.

iii. 차분

차분을 이해하기 위해서, 후향연산자(Backshift Operator)에 대해 먼저 알아보겠습니다. 차분에서 사용되는 후향연산자는 바로 한 시점 전으로 돌려주는 작용을 하는 연산자입니다. 다음과 같이 표현할 수 있습니다.

$$BX_t = X_{t-1}$$

후향연산자를 이해했다면, 이제 차분이 무엇인지 알아보겠습니다. 차분은 관측값들의 차이를 구하는 것입니다. 즉, 차이를 통해 추세와 계절성을 제거할 수 있는 방법입니다. 후향연산자를 사용하여 다음과 같이 표현할 수 있습니다.

[1차 차분]

$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t$$

[2차 차분]

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2} = (1-B)^2 X_t$$

- **A [추세만 존재하는 경우] Differencing**

추세를 $m_t = (c_0 + c_1 t)$인 선형이라고 가정했을 때, 차분을 진행해보겠습니다.

$$\nabla m_t = m_t - m_{t-1} = (c_0 + c_1 t) - (c_0 + c_1(t-1)) = c_1$$

위와 같이, 시간 t에 영향을 받지 않는 상수 c_1만 남아 추세가 제거되었음을 확인할 수 있습니다. 일반적으로 k차 차분을 하면, k차 추세 (k-th order polynomial trend)를 제거할 수 있습니다.

- **B [계절성만 존재하는 경우] Seasonal Differencing**

계절성이 존재하는 경우에는 lag-d differencing을 통해 계절성을 제거합니다. lag-d difference 연산자는 다음과 같습니다.

$$\nabla_d X_t = (1 - B^d) X_t$$

이때, $s_t = s_{t+d}$를 가정한 뒤 계절성이 존재하는 시계열에 lag-d 차분을 적용합니다. 이후 오차항만 남아 계절성이 제거됩니다.

> 🥦 **d차 차분과 lag-d 차분의 차이점**
>
> d차 차분은 $\nabla^d = (1-B)^d$로, lag-d 차분은 $\nabla_d = (1-B^d)$로 표현합니다.
> d차 차분은 1차 차분을 진행한 뒤, 차분을 한 번 더 하는 것이라면, lag-d 차분은 시점 t와 $t-d$의 값의 차이를 계산하는 것입니다.

IV. 정상성 검정

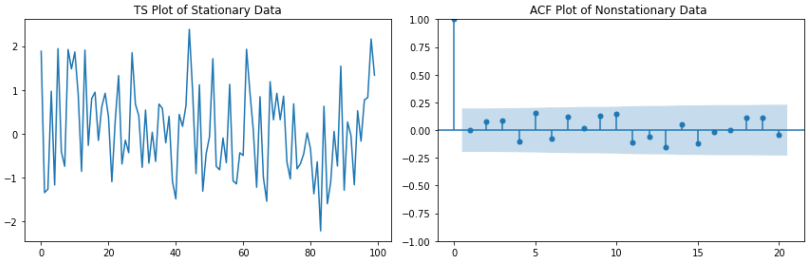
위와 같은 방법들을 통해, 시계열 자료에서 비정상 부분을 제거하였다면 **정상성을 만족하는 오차**만 남아야 합니다. 지금부터 남아 있는 오차들이 정상성을 만족하는지 확인하는 정상성 검정 과정에 대해 알아보겠습니다.

1. 자기공분산함수(ACVF)와 자기상관함수(ACF)

정상성을 만족하지 않는다는 것은 곧 확률적 성질이 시간에 의존한다는 것입니다. 정상성을 확인하기 위하여 평균, 분산 뿐만 아니라 시간에 따른 상관 정도(시계열의 확률적 성질이 시간 t에 의존하는 정도)를 나타내기 위한 **자기공분산함수(autocovariance function)와 자기상관함수(autocorrelation function)**에 대해 알아보겠습니다.

자기공분산함수 ACVF (Autocovariance Function)	**자기상관함수 ACF** (Autocorrelation Function)
$\gamma_k = cov(X_t, X_{t+k})$	$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$
표본자기공분산함수 SACVF (Sample Autocovariance Function)	**표본자기상관함수 SACF** (Sample Autocorrealtion Function)
$\hat{\gamma_X}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(X_j - \bar{X})(X_{j+h} - \bar{X})$	$\hat{\rho_X}(h) = \frac{\hat{\gamma_X}(h)}{\gamma_X(0)}, , \hat{\rho}(0) = 1$

일반적으로 자기상관함수 플랏을 통해 정상성 검정을 진행합니다.



위와 같은 stationary한 데이터에 대한 ACF plot은 시차(lag) 간의 자기상관이 0 이후에서 급격히 감소합니다. 데이터 간의 의존성이나 상관성이 일정하지 않고, 특정 시차 이후에는 거의 사라지는 것 역시 확인할 수 있습니다.

2. 백색잡음(White Noise)

자기상관이 없는 시계열을 백색잡음이라고 합니다. 백색잡음은 대표적인 정상 확률과정으로 $\{X_t\}$는 상관관계가 존재하지 않고(uncorrelated), 평균이 0, 분산이 $\sigma^2 < \infty$이면 백색잡음이라고 부르고, 다음과 같이 표기합니다.

$$X_t \sim WN(0, \sigma^2)$$

*$IID(0, \sigma^2)$는 백색잡음이라고 할 수 있지만, 그 역은 true가 아니라는 점 주의하세요!

3. 백색잡음 검정

비정상 시계열로부터 정상적으로 추세와 계절성을 제거했다면, 남아 있는 오차항은 IID 혹은 WN 조건을 만족합니다. 이 경우 우리는 σ^2을 구하기 위해 $\gamma_X(0)$만 추정해주면 됩니다. 아래의 과정을 통해 백색잡음 검정 방법을 알아보겠습니다. 백색잡음 검정은 자기상관, 정규성, 정상성에 대한 검정을 진행합니다.

i. 자기상관 검정

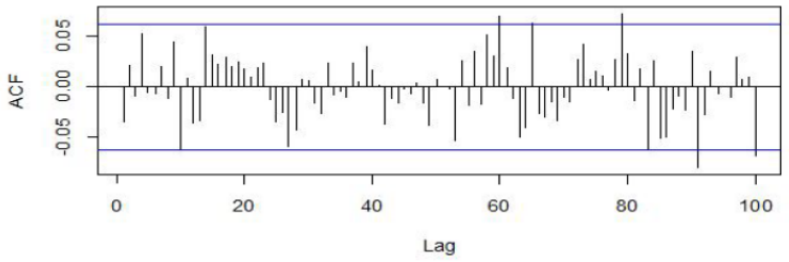
$$\hat{\rho} \approx N(0, \frac{1}{n})$$

오차가 백색잡음 $WN(0,1)$을 따른다면, 표본자기상관함수 $\rho(\hat{h})$는 평균이 0이고 분산이 $\frac{1}{n}$인 정규분포로 근사합니다. 이러한 사실을 바탕으로 다음의 가설 검정을 진행합니다.

$$H_0 : \rho(h) = 0 \; vs \; H_1 : \rho(h) \neq 0$$

귀무가설: 자기상관이 존재하지 않는다. vs 대립가설: 자기상관이 존재한다.

만약 $|\rho(\hat{h})|$가 $1.96/\sqrt{n}$안에 존재한다면귀무가설을 기각할 수 없습니다. 즉, 오차항에 자기상관이 존재하지 않는다고 판단합니다. 검정 결과는 ACF Plot(=correlogram)을 통해 시각적으로도 확인할 수 있습니다.



그래프의 x축은 시차, y축은 acf를 의미하여 파란선을 통해 신뢰구간을 확인할 수 있습니다. 위 그림에서 대부분의 경우 신뢰구간을 벗어나지 않기 때문에 오차항에 자기상관이 존재하지 않는다고 판단합니다.

이 외에도 portmanteau test, Ljung-Box test, McLeod-Li test 등을 이용하여 확인할 수 있습니다

ii. 정규성 검정

H_0 : 정규성이 존재한다. vs H_1 : 정규성이 존재하지 않는다.

	특징
QQ plot	시각적으로 확인할 수 있는 방법
KS test	표본의 누적확률분포가 모집단의 누적확률분포와 얼마나 유사한지 비교하는 방법
Jarque-Bera test	왜도와 첨도를 통해 정규성을 검정하는 방법

iii. 정상성 검정

	특징	H_0
Kpss test	단위근 검정방법 중 하나	정상 시계열이다.
ADF test	단위근 검정방법 중 하나	정상 시계열이다.
PP test	이분산이 있는 경우에도 사용 가능한 검정 방법	정상 시계열이다.

Appendix

[Harmonic Regression에서 λ_j와 k의 선택]

$$s_t = a_0 + \sum_{j=1}^{k}(a_j)cos(\lambda_j t) + b_j sin(\lambda_j t))$$

λ_j는 주기가 2π인 함수의 주기와 데이터의 주기를 맞춰 주기 위한 값입니다.

- 주기 반복 횟수: $f_1 = n/d$ (n=데이터 개수, d=주기)

  → $f_j = jf_1$
- $\lambda_j = f_j \times (2\pi/n)$

k는 주로 1~4 사이의 값을 사용합니다.

ex. n=72, d=12

→ $f_1 = 72/12 = 6$

$\lambda_j = j \times 6 \times (2\pi/72)$

[이동평균 평활법(Moving Average Smoothing)에서 구간 평균이 추세와 같아지는 이유]

[1] 길이가 2q+1인 구간의 평균은 다음과 같습니다.

$$\sum_{j=-q}^{j=q}(m_{t+j} + Y_{t+j}) = \frac{1}{2q+1}\sum_{j=-q}^{j=q} m_{t+j} + \frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j}$$

[2] 위 식에 추세성분 m_t를 대입합니다.

$$m_t = c_0 + c_1 t,\ E(Y_t) = 0$$

$$\frac{1}{2q+1}\sum_{j=-q}^{j=q} m_{t+j} = c_0 + c_1 t = m_t, t \in [q+1, n-q]$$

$$\frac{1}{2q+1}\sum_{j=-q}^{j=q} Y_{t+j} \approx E(Y_t) = 0 \text{ (by WLLN)}$$